PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-372987
(43)Date of publication of application : 26.12.2002

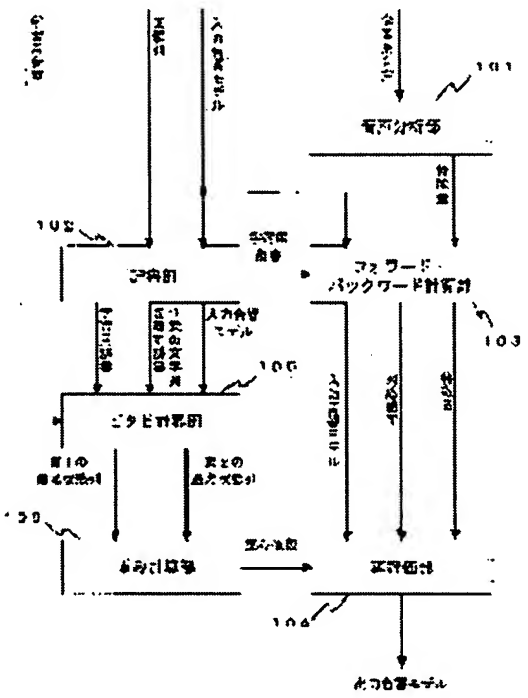
(51)Int.Cl. G10L 15/14
G10L 15/06

(21)Application number : 2001-179125
(22)Date of filing : 13.06.2001

(71)Applicant : NEC CORP
(72)Inventor : TAKANO MASARU

(54) ACOUSTIC MODEL LEARNING DEVICE, ACOUSTIC MODEL LEARNING METHOD, AND PROGRAM FOR THE SAME

(57)Abstract:
PROBLEM TO BE SOLVED: To provide an acoustic model learning device, an acoustic model learning method and a program for the same for extracting only a voice sample useful for generating an acoustic model out of observed voice samples to generate a highly reliable acoustic model.
SOLUTION: The reevaluation section 104 calculates a statistical quantity based on the feature amount of the voice for learning extracted by a voice analyzing section 104 and correspondence probability calculated by a forward/ backward calculating section 103 and a weight coefficient Rt calculated by a weight calculating section 106 to re-estimate the acoustic model and output the output acoustic model.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報（A）**　(11)特許出願公開番号
特開2002－372987
（P2002－372987A）
(43)公開日　平成14年12月26日(2002.12.26)

(51)Int.Cl.[7]	識別記号	ＦＩ	テーマコード(参考)
Ｇ１０Ｌ　15/14 15/06		Ｇ１０Ｌ　3/00	５３５Ｚ　５Ｄ０１５ ５２１Ｆ

審査請求　未請求　請求項の数18　ＯＬ　（全 17 頁）

(21)出願番号　特願2001－179125(P2001－179125)

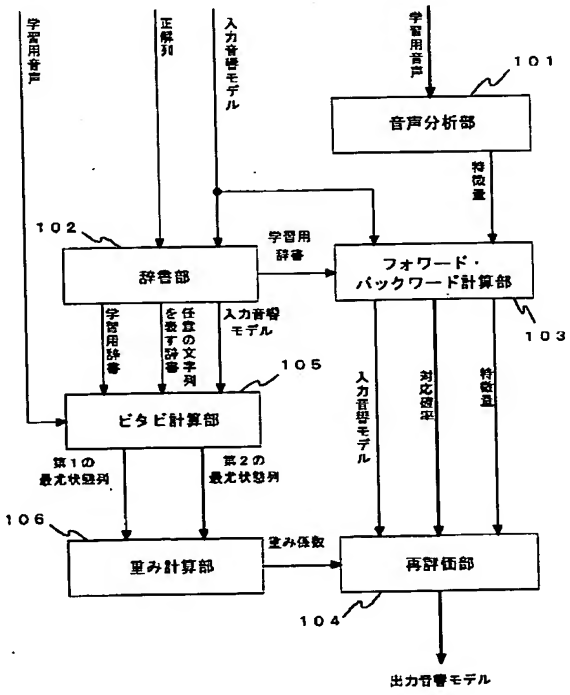
(22)出願日　平成13年６月13日(2001.6.13)

(71)出願人　000004237
日本電気株式会社
東京都港区芝五丁目７番１号
(72)発明者　高野　優
東京都港区芝五丁目７番１号　日本電気株式会社内
(74)代理人　100084250
弁理士　丸山　隆夫

Ｆターム(参考)　5D015 GG00 HH00

(54)【発明の名称】　音響モデル学習装置、音響モデル学習方法、およびそのプログラム

(57)【要約】
【課題】　観測された音声サンプルのうち音響モデルの作成に有用なものだけを抽出し、信頼性の高い音響モデルを作成する音響モデル学習装置、音響モデル学習方法、およびプログラムを提供する。
【解決手段】　再評価部１０４は、音声分析部１０１により抽出された学習用音声の特徴量と、フォワード・バックワード計算部１０３により算出された対応確率と、重み計算部１０６により算出された重み係数R_tと、に基づいて統計量を算出し、音響モデルの再推定を行い、出力音響モデルを出力する。



【特許請求の範囲】
【請求項１】　入力される学習用音声からフレームごとに特徴量を抽出する音声分析手段と、
所定の音声からフレームごとに抽出された特徴量を示す確率分布を用いて、前記所定の音声のフレームごとに分割された前記所定の音声の断片を状態として表現し、該状態を構成単位とする入力音響モデルと、前記学習用音声の内容を示す文字列情報である正解列と、に基づいて、前記入力音響モデルにおける前記状態に前記正解列を割り当てた状態列の情報である学習用辞書を生成する辞書生成手段と、
該辞書生成手段により生成された学習用辞書を参照し、前記学習用音声の特徴量と前記入力音響モデルにおける状態との対応確率を前記学習用音声のフレームごとに算出する対応確率算出手段と、
所定の文字列を用いて、前記入力音響モデルにより表現される前記状態あるいは複数の前記状態からなる状態列を、前記学習用音声のフレームごとに最尤に割り当て、所定の最尤状態列を生成する最尤状態列生成手段と、
該最尤状態列生成手段により生成された所定の最尤状態列に基づいて、前記対応確率に重み付けする際に付加する係数である重み係数を、前記学習用音声のフレームごとに算出する重み計算手段と、
前記対応確率算出手段により算出された対応確率と、前記重み計算手段により算出された重み係数と、前記音声分析手段により算出された特徴量と、に基づいて統計量を算出し、該算出した統計量に基づいて、前記入力音響モデルのパラメータを再推定し、出力音響モデルを作成する再評価手段と、
を有することを特徴とする音響モデル学習装置。
【請求項２】　前記再評価手段は、
前記学習用音声のフレームごとの前記対応確率に、前記重み係数を乗算し、前記学習用音声のフレームごとの対応確率に重み付けを行い、該重み付けされた対応確率を用いて前記統計量を算出し、該算出した統計量に基づいて、前記入力音響モデルのパラメータを再推定し、前記出力音響モデルを作成することを特徴とする請求項１記載の音響モデル学習装置。
【請求項３】　前記重み計算手段は、前記最尤状態列生成手段により、前記学習用辞書を用いて生成された最尤状態列を第１の最尤状態列とし、任意の文字列を用いて生成された最尤状態列を第２の最尤状態列とした場合、前記学習用音声のフレームごとに、前記第１の最尤状態列と前記第２の最尤状態列とを比較し、該比較に基づいて、前記学習用音声のフレームごとに前記重み係数を算出することを特徴とする請求項１または２記載の音響モデル学習装置。
【請求項４】　前記重み計算手段は、
前記学習用音声のフレームごとに、前記第１の最尤状態列と前記第２の最尤状態列とを比較し、前記割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは前記重み係数を１とし、互いに異なるフレームでは前記重み係数を１より小さな値として算出することを特徴とする請求項３記載の音響モデル学習装置。
【請求項５】　前記重み計算手段は、
前記学習用音声のフレームごとに、前記第１の最尤状態列と前記第２の最尤状態列とを比較し、前記割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは前記重み係数を１とし、互いに異なるフレームでは前記重み係数を１より大きな値として算出することを特徴とする請求項３記載の音響モデル学習装置。
【請求項６】　前記重み計算手段は、
前記割り当てられた状態ごとに、算出した前記重み係数の和をそれぞれ算出し、該算出した重み係数の和が、それぞれ等しい値となるように前記算出した重み係数を設定することを特徴とする請求項１から５のいずれか１項に記載の音響モデル学習装置。
【請求項７】　入力される学習用音声からフレームごとに特徴量を抽出する音声分析工程と、
所定の音声からフレームごとに抽出された特徴量を示す確率分布を用いて、前記所定の音声におけるフレームごとに分割された前記所定の音声の断片を状態として表現し、該状態を構成単位とする入力音響モデルと、前記学習用音声の内容を示す文字列情報である正解列と、に基づいて、前記入力音響モデルにおける前記状態に前記正解列を割り当てた状態列の情報である学習用辞書を生成する辞書生成工程と、
該辞書生成工程により生成された学習用辞書を参照し、前記学習用音声の特徴量と前記入力音響モデルにおける状態との対応確率を前記学習用音声のフレームごとに算出する対応確率算出工程と、
所定の文字列を用いて、前記入力音響モデルにより表現される前記状態あるいは複数の前記状態からなる状態列を、前記学習用音声のフレームごとに最尤に割り当て、所定の最尤状態列を生成する最尤状態列生成工程と、
該最尤状態列生成工程により生成された所定の最尤状態列に基づいて、前記対応確率に重み付けする際に付加する係数である重み係数を、前記学習用音声のフレームごとに算出する重み計算工程と、
前記対応確率算出工程により算出された対応確率と、前記重み計算工程により算出された重み係数と、前記音声分析工程により算出された特徴量と、に基づいて統計量を算出し、該算出した統計量に基づいて、前記入力音響モデルのパラメータを再推定し、出力音響モデルを作成する再評価工程と、
を有することを特徴とする音響モデル学習方法。
【請求項８】　前記再評価工程は、
前記学習用音声のフレームごとの前記対応確率に、前記重み係数を乗算し、前記学習用音声のフレームごとの対応確率に重み付けを行い、該重み付けされた対応確率を

用いて前記統計量を算出し、該算出した統計量に基づいて、前記入力音響モデルのパラメータを再推定し、前記出力音響モデルを作成することを特徴とする請求項７記載の音響モデル学習方法。

【請求項９】　前記重み計算工程は、

前記最尤状態列生成工程により、前記学習用辞書を用いて生成された最尤状態列を第１の最尤状態列とし、任意の文字列を用いて生成された最尤状態列を第２の最尤状態列とした場合、

前記学習用音声のフレームごとに、前記第１の最尤状態列と前記第２の最尤状態列とを比較し、該比較に基づいて、前記学習用音声のフレームごとに前記重み係数を算出することを特徴とする請求項７または８記載の音響モデル学習方法。

【請求項１０】　前記重み計算工程は、

前記学習用音声のフレームごとに、前記第１の最尤状態列と前記第２の最尤状態列とを比較し、前記割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは前記重み係数を１とし、互いに異なるフレームでは前記重み係数を１より小さな値として算出することを特徴とする請求項９記載の音響モデル学習方法。

【請求項１１】　前記重み計算工程は、

前記学習用音声のフレームごとに、前記第１の最尤状態列と前記第２の最尤状態列とを比較し、前記割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは前記重み係数を１とし、互いに異なるフレームでは前記重み係数を１より大きな値として算出することを特徴とする請求項９記載の音響モデル学習方法。

【請求項１２】　前記重み計算工程は、

前記割り当てられた状態ごとに、算出した前記重み係数の和をそれぞれ算出し、該算出した重み係数の和が、それぞれ等しい値となるように前記算出した重み係数を設定することを特徴とする請求項７から１１のいずれか１項に記載の音響モデル学習方法。

【請求項１３】　入力される学習用音声からフレームごとに特徴量を抽出する音声分析処理と、

所定の音声からフレームごとに抽出された特徴量を示す確率分布を用いて、前記所定の音声におけるフレームごとに分割された前記所定の音声の断片を状態として表現し、該状態を構成単位とする入力音響モデルと、前記学習用音声の内容を示す文字列情報である正解列と、に基づいて、前記入力音響モデルにおける前記状態に前記正解列を割り当てた状態列の情報である学習用辞書を生成する辞書生成処理と、

該辞書生成処理により生成された学習用辞書を参照し、前記学習用音声の特徴量と前記入力音響モデルにおける状態との対応確率を前記学習用音声のフレームごとに算出する対応確率算出処理と、

所定の文字列を用いて、前記入力音響モデルにより表現される前記状態あるいは複数の前記状態からなる状態列を、前記学習用音声のフレームごとに最尤に割り当て、所定の最尤状態列を生成する最尤状態列生成処理と、

該最尤状態列生成処理により生成された所定の最尤状態列に基づいて、前記対応確率に重み付けする際に付加する係数である重み係数を、前記学習用音声のフレームごとに算出する重み計算処理と、

前記対応確率算出処理により算出された対応確率と、前記重み計算処理により算出された重み係数と、前記音声分析処理により算出された特徴量と、に基づいて統計量を算出し、該算出した統計量に基づいて、前記入力音響モデルのパラメータを再推定し、出力音響モデルを作成する再評価処理と、

を実行させるためのプログラム。

【請求項１４】　前記再評価処理は、

前記学習用音声のフレームごとの前記対応確率に、前記重み係数を乗算し、前記学習用音声のフレームごとの対応確率に重み付けを行い、該重み付けされた対応確率を用いて前記統計量を算出し、該算出した統計量に基づいて、前記入力音響モデルのパラメータを再推定し、前記出力音響モデルを作成することを特徴とする請求項１３記載のプログラム。

【請求項１５】　前記重み計算処理は、

前記最尤状態列生成処理により、前記学習用辞書を用いて生成された最尤状態列を第１の最尤状態列とし、任意の文字列を用いて生成された最尤状態列を第２の最尤状態列とした場合、

前記学習用音声のフレームごとに、前記第１の最尤状態列と前記第２の最尤状態列とを比較し、該比較に基づいて、前記学習用音声のフレームごとに前記重み係数を算出することを特徴とする請求項１３または１４記載のプログラム。

【請求項１６】　前記重み計算処理は、

前記学習用音声のフレームごとに、前記第１の最尤状態列と前記第２の最尤状態列とを比較し、前記割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは前記重み係数を１とし、互いに異なるフレームでは前記重み係数を１より小さな値として算出することを特徴とする請求項１５記載のプログラム。

【請求項１７】　前記重み計算処理は、

前記学習用音声のフレームごとに、前記第１の最尤状態列と前記第２の最尤状態列とを比較し、前記割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは前記重み係数を１とし、互いに異なるフレームでは前記重み係数を１より大きな値として算出することを特徴とする請求項１５記載の音響モデル学習装置。

【請求項１８】　前記重み計算処理は、

前記割り当てられた状態ごとに、算出した前記重み係数の和をそれぞれ算出し、該算出した重み係数の和が、それぞれ等しい値となるように前記算出した重み係数を設

定することを特徴とする請求項１３から１７のいずれか１項に記載の音響モデル学習装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、音響モデル学習装置、音響モデル学習方法、およびそのプログラムに関し、特に、音声サンプルの特性に応じて音声サンプルに重み付けを行い、信頼性の高い音響モデルを作成する音響モデル学習装置、音響モデル学習方法、およびそのプログラムに関する。

【０００２】

【従来の技術】音響モデル学習装置は、実際の音声を用いて、音声認識に使用される音響モデルを学習することが多い。一般に、学習される音響モデルとして、ＨｉｄｄｅｎＭａｒｋｏｖ　Ｍｏｄｅｌ（隠れマルコフモデル、以下、HMMとする）が用いられる。また、HMMにおける状態を表す確率分布としては、連続混合分布が用いられる場合が多い。また、多くの場合、HMMの学習には、フォワード・バックワード法が用いられる。上記のようなHMMによる音響モデルのパラメータの推定について記載されている文献としては、Ｌａｗｒａｎｃｅ Ｌａｂｉｎｅｒ，Ｂｉｉｎｇ－Ｈｗａｎｇ Ｊｕａｎｇ「Ｆｕｎｄａｍｅｎｔａｌｓ ｏｆＳｐｅｅｃｈ Ｒｅｃｏｇｎｉｔｉｏｎ １９９３ ｐ．３３３～ｐ．３８９」（以下、従来例１）があった。

【０００３】従来例１では、HMMに用いられる連続混合確率分布を構成する複数の確率分布それぞれに、連続混合確率分布における混合比を示す混合重みを付加していた。

【０００４】以下、フォワード・バックワード法を用いたHMMにおけるパラメータの計算方法について説明する。

【０００５】時刻（フレーム）ｔごとの特徴量をO_t（ｔは1以上T以下の整数）とすると、フォワード・バックワード法におけるフォワード確率αは、以下に示す（式１．１）および（式１．２）により示される。

【０００６】

【数１】

$$\alpha(1,i)=b(i,O_1) \quad \ldots \quad (式1.1)$$

$$\alpha(t+1,j)=\left[\sum_{i=1}^{N}\alpha(t,i)a_{ij}\right]b(j,O_{t+1}) \quad \ldots \quad (式1.2)$$

【０００７】なお、フォワード確率α（ｔ，ｉ）は、特徴量O_tを観測し、状態S_iにある確率を示す。同様に、フォワード確率α（１，ｉ）は、特徴量O_1を観測し、状態S_iにある確率、フォワード確率α（ｔ＋１，ｊ）は、特徴量O_{t+1}を観測し、状態S_jにある確率を示す。

【０００８】また、状態遷移確率a_{ij}は、状態S_iから状態S_jへ遷移する確率を表す。観測確率ｂ（ｉ，O_1）は、状態S_iに遷移する際に、フレームｔにおける特徴量O_tが観測される確率を示す。

【０００９】また、フォワード・バックワード法におけるバックワード確率βは、以下に示す（式２．１）および（式２．２）により示される。

【００１０】

【数２】

$$\beta(T,i)=1 \quad \ldots \quad (式2.1)$$

$$\beta(t,i)=\sum_{j=1}^{N}a_{ij}\,b(j,O_{t+1})\,\beta(t+1,j) \quad \ldots \quad (式2.2)$$

【００１１】なお、バックワード確率β（ｔ，ｉ）は、フレームｔにおいて状態S_iにあり、以後フレーム（ｔ＋１）において特徴量O_{t+1}を観測する確率を示す。フレームTは、最終状態におけるフレームを表す。

【００１２】また、フォワード・バックワード法における対応確率γは、フォワード確率αとバックワード確率βとに基づいて、計算される。対応確率γは、以下に示す（式３．１）により示される。

【００１３】

【数３】

$$\gamma(t,j,k)=\frac{\alpha(t,j)\beta(t,j)}{N}\cdot\frac{c_{jk}N(O_t,\mu_{jk},U_{jk})}{}$$

クトルμ（t，j，k）、および共分散行列U（j，k）の各平均は、以下に示す（式4．1）、（式4．2）、および（式4．3）により計算される。

【0016】

【数4】

$$\overline{c}_{jk}=\frac{\sum_{t=1}^{T}\gamma(t,j,k)}{\sum_{s.t.\,Ot=Vk}\sum_{t=1}^{T}\sum_{k=1}^{B}\gamma(t,j,k)} \quad \cdots \quad (式4.1)$$

$$\overline{\mu}(t,j,k)=\frac{\sum_{t=1}^{T}\gamma(t,j,k)\cdot Ot}{\sum_{s.t.\,Ot=Vk}\sum_{t=1}^{T}\gamma(t,j,k)} \quad \cdots \quad (式4.2)$$

$$\overline{U}(j,k)=\frac{\sum_{t=1}^{T}\gamma(t,j,k)\cdot(Ot-\mu_{jk})(Ot-\mu_{jk})'}{\sum_{s.t.\,Ot=Vk}\sum_{t=1}^{T}\gamma(t,j,k)} \quad \cdots \quad (式4.3)$$

【0017】なお、混合重みc_{jk}は、HMMにおける状態S_jのk番目の混合分布要素に対する混合重みである。また、平均ベクトルμ（t，j，k）は、HMMにおける状態S_jのk番目の混合分布要素の平均ベクトルである。また、共分散行列U（j，k）は、HMMにおける状態S_jのk番目の混合分布要素の共分散行列である。また、V_kは、文字列Vにおける所定の文字を示す。また、$(O_t-\mu_{jk})'$は、ベクトル$(O_t-\mu_{jk})$の対置ベクトルを表す。

【0018】また、特開平5－232989号公報が開示するところの音響モデルの話者適応化法（以下、従来例2）では、HMMに用いられる連続混合確率分布を構成する複数の確率分布それぞれの混合比を決める重み係数だけを再推定していた。

【0019】また、特開平10－11086号公報が開示するところの隠れマルコフモデルの計算方式（以下、従来例3）には、フォワードバックワード法を用いたHMMの計算方式が記載されていた。

【0020】

【発明が解決しようとする課題】一般に、信頼性の高い確率モデルの学習には、大量の音声データが必要となる。特に、不特定話者用の音響モデルには、話者の個人差による音声の変動を吸収する必要がある。従って、不特定話者用の音響モデルには、話者の発声による音声データが多数必要となる。しかしながら、大量の音声サンプルを収集する際には、話者の誤発声あるいは低品質の音声が混入する可能性がある。

【0021】さらに、確率モデル（音響モデル）の推定を行う場合に、以下に示すような問題が生じてしまう。通常、音声データを収集する際、話者の自然な発声による音声データを得る必要がある。従って、音声データとして収集される話者の発声内容は、実際に存在する単語が用いられる。また、実際に存在する単語を構成する音素（文字）の分布には必然的に偏りが生じる。例えば、日本語の場合は、母音、特に「あ」の出現頻度が非常に高い。確率モデルを推定する場合、確率分布を推定するサンプル数によって確率分布の信頼性に格差が生じてしまう。従って、単語を構成する音素を音響モデルを構築する音声データとして用いる場合、音素の出現頻度の偏りを修正する必要がある。

【0022】本発明は、上記問題点に鑑みてなされたものであり、従来例1、従来例2、および従来例2と従来例3とを組み合わせたものにおいてHMMの各混合分布要素に付加されている重みに加え、収集した音声サンプルの特性に応じて設定された重み係数を、音声サンプルの各フレームにさらに付加することによって、特定の音声サンプルあるいは音声サンプルの特定部分を音響モデルの学習の際に増幅あるいは除去し、音声サンプルを構成する音素の出現頻度の偏りを修正し、信頼性の高い音響モデルを提供する音響モデル学習装置を提供することを目的とする。

【0023】

【課題を解決するための手段】かかる目的を達成するため、請求項１記載の発明は、入力される学習用音声からフレームごとに特徴量を抽出する音声分析手段と、所定の音声からフレームごとに抽出された特徴量を示す確率分布を用いて、所定の音声におけるフレームごとに分割された所定の音声の断片を状態として表現し、状態を構成単位とする入力音響モデルと、学習用音声の内容を示す文字列情報である正解列と、に基づいて、入力音響モデルにおける状態に正解列を割り当てた状態列の情報である学習用辞書を生成する辞書生成手段と、辞書生成手段により生成された学習用辞書を参照し、学習用音声の特徴量と入力音響モデルにおける状態との対応確率を学習用音声のフレームごとに算出する対応確率算出手段と、所定の文字列を用いて、入力音響モデルにより表現される状態あるいは複数の状態からなる状態列を、学習用音声のフレームごとに最尤に割り当て、所定の最尤状態列を生成する最尤状態列生成手段と、最尤状態列生成手段により生成された所定の最尤状態列に基づいて、対応確率に重み付けする際に付加する係数である重み係数を、学習用音声のフレームごとに算出する重み計算手段と、対応確率算出手段により算出された対応確率と、重み計算手段により算出された重み係数と、音声分析手段により算出された特徴量と、に基づいて統計量を算出し、算出した統計量に基づいて、入力音響モデルのパラメータを再推定し、出力音響モデルを作成する再評価手段と、を有することを特徴とする。

【００２４】また、請求項２記載の発明によれば、請求項１記載の音響モデル学習装置において、再評価手段は、学習用音声のフレームごとの対応確率に、重み係数を乗算し、学習用音声のフレームごとの対応確率に重み付けを行い、重み付けされた対応確率を用いて統計量を算出し、算出した統計量に基づいて、入力音響モデルのパラメータを再推定し、出力音響モデルを作成することを特徴とする。

【００２５】また、請求項３記載の発明によれば、請求項１または２記載の音響モデル学習装置において、重み計算手段は、最尤状態列生成手段により、学習用辞書を用いて生成された最尤状態列を第１の最尤状態列とし、任意の文字列を用いて生成された最尤状態列を第２の最尤状態列とした場合、学習用音声のフレームごとに、第１の最尤状態列と第２の最尤状態列とを比較し、比較に基づいて、学習用音声のフレームごとに重み係数を算出することを特徴とする。

【００２６】また、請求項４記載の発明によれば、請求項３記載の音響モデル学習装置において、重み計算手段は、学習用音声のフレームごとに、第１の最尤状態列と第２の最尤状態列とを比較し、割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは重み係数を１とし、互いに異なるフレームでは重み係数を１より小さな値として算出することを特徴とする。

【００２７】また、請求項５記載の発明によれば、請求項３記載の音響モデル学習装置において、重み計算手段は、学習用音声のフレームごとに、第１の最尤状態列と第２の最尤状態列とを比較し、割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは重み係数を１とし、互いに異なるフレームでは重み係数を１より大きな値として算出することを特徴とする。

【００２８】また、請求項６記載の発明によれば、請求項１から５のいずれか１項に記載の音響モデル学習装置において、重み計算手段は、割り当てられた状態ごとに、算出した重み係数の和をそれぞれ算出し、算出した重み係数の和が、それぞれ等しい値となるように算出した重み係数を設定することを特徴とする。

【００２９】また、請求項７記載の発明は、入力される学習用音声からフレームごとに特徴量を抽出する音声分析工程と、所定の音声からフレームごとに抽出された特徴量を示す確率分布を用いて、所定の音声におけるフレームごとに分割された所定の音声の断片を状態として表現し、状態を構成単位とする入力音響モデルと、学習用音声の内容を示す文字列情報である正解列と、に基づいて、入力音響モデルにおける状態に正解列を割り当てた状態列の情報である学習用辞書を生成する辞書生成工程と、辞書生成工程により生成された学習用辞書を参照し、学習用音声の特徴量と入力音響モデルにおける状態との対応確率を学習用音声のフレームごとに算出する対応確率算出工程と、所定の文字列を用いて、入力音響モデルにより表現される状態あるいは複数の状態からなる状態列を、学習用音声のフレームごとに最尤に割り当て、所定の最尤状態列を生成する最尤状態列生成工程と、最尤状態列生成工程により生成された所定の最尤状態列に基づいて、対応確率に重み付けする際に付加する係数である重み係数を、学習用音声のフレームごとに算出する重み計算工程と、対応確率算出工程により算出された対応確率と、重み計算工程により算出された重み係数と、音声分析工程により算出された特徴量と、に基づいて統計量を算出し、算出した統計量に基づいて、入力音響モデルのパラメータを再推定し、出力音響モデルを作成する再評価工程と、を有することを特徴とする音響モデル学習方法。

【００３０】また、請求項８記載の発明によれば、請求項７記載の音響モデル学習方法において、再評価工程は、学習用音声のフレームごとの対応確率に、重み係数を乗算し、学習用音声のフレームごとの対応確率に重み付けを行い、重み付けされた対応確率を用いて統計量を算出し、算出した統計量に基づいて、入力音響モデルのパラメータを再推定し、出力音響モデルを作成することを特徴とする。

【００３１】また、請求項９記載の発明によれば、請求項７または８記載の音響モデル学習方法において、重み計算工程は、最尤状態列生成工程により、学習用辞書を

用いて生成された最尤状態列を第１の最尤状態列とし、任意の文字列を用いて生成された最尤状態列を第２の最尤状態列とした場合、学習用音声のフレームごとに、第１の最尤状態列と第２の最尤状態列とを比較し、比較に基づいて、学習用音声のフレームごとに重み係数を算出することを特徴とする。

【００３２】また、請求項１０記載の発明によれば、請求項９記載の音響モデル学習方法において、重み計算工程は、学習用音声のフレームごとに、第１の最尤状態列と第２の最尤状態列とを比較し、割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは重み係数を１とし、互いに異なるフレームでは重み係数を１より小さな値として算出することを特徴とする。

【００３３】また、請求項１１記載の発明によれば、請求項９記載の音響モデル学習方法において、重み計算工程は、学習用音声のフレームごとに、第１の最尤状態列と第２の最尤状態列とを比較し、割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは重み係数を１とし、互いに異なるフレームでは重み係数を１より大きな値として算出することを特徴とする。

【００３４】また、請求項１２記載の発明によれば、請求項７から１１のいずれか１項に記載の音響モデル学習方法において、重み計算工程は、割り当てられた状態ごとに、算出した重み係数の和をそれぞれ算出し、算出した重み係数の和が、それぞれ等しい値となるように算出した重み係数を設定することを特徴とする。

【００３５】また、請求項１３記載の発明は、入力される学習用音声からフレームごとに特徴量を抽出する音声分析処理と、所定の音声からフレームごとに抽出された特徴量を示す確率分布を用いて、所定の音声におけるフレームごとに分割された所定の音声の断片を状態として表現し、状態を構成単位とする入力音響モデルと、学習用音声の内容を示す文字列情報である正解列と、に基づいて、入力音響モデルにおける状態に正解列を割り当てた状態列の情報である学習用辞書を生成する辞書生成処理と、辞書生成処理により生成された学習用辞書を参照し、学習用音声の特徴量と入力音響モデルにおける状態との対応確率を学習用音声のフレームごとに算出する対応確率算出処理と、所定の文字列を用いて、入力音響モデルにより表現される状態あるいは複数の状態からなる状態列を、学習用音声のフレームごとに最尤に割り当て、所定の最尤状態列を生成する最尤状態列生成処理と、最尤状態列生成処理により生成された所定の最尤状態列に基づいて、対応確率に重み付けする際に付加する係数である重み係数を、学習用音声のフレームごとに算出する重み計算処理と、対応確率算出処理により算出された対応確率と、重み計算処理により算出された重み係数と、音声分析処理により算出された特徴量と、に基づいて統計量を算出し、算出した統計量に基づいて、入力音響モデルのパラメータを再推定し、出力音響モデルを作成する再評価処理と、を実行させることを特徴とする。

【００３６】また、請求項１４記載の発明によれば、請求項１３記載のプログラムにおいて、再評価処理は、学習用音声のフレームごとの対応確率に、重み係数を乗算し、学習用音声のフレームごとの対応確率に重み付けを行い、重み付けされた対応確率を用いて統計量を算出し、算出した統計量に基づいて、入力音響モデルのパラメータを再推定し、出力音響モデルを作成することを特徴とする。

【００３７】また、請求項１５記載の発明によれば、請求項１３または１４記載のプログラムにおいて、重み計算処理は、最尤状態列生成処理により、学習用辞書を用いて生成された最尤状態列を第１の最尤状態列とし、任意の文字列を用いて生成された最尤状態列を第２の最尤状態列とした場合、学習用音声のフレームごとに、第１の最尤状態列と第２の最尤状態列とを比較し、比較に基づいて、学習用音声のフレームごとに重み係数を算出することを特徴とする。

【００３８】また、請求項１６記載の発明によれば、請求項１５記載のプログラムにおいて、重み計算処理は、学習用音声のフレームごとに、第１の最尤状態列と第２の最尤状態列とを比較し、割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは重み係数を１とし、互いに異なるフレームでは重み係数を１より小さな値として算出することを特徴とする。

【００３９】また、請求項１７記載の発明によれば、請求項１５記載の音響モデル学習装置において、重み計算処理は、学習用音声のフレームごとに、第１の最尤状態列と第２の最尤状態列とを比較し、割り当てられた状態あるいは複数の状態からなる状態列が一致したフレームでは重み係数を１とし、互いに異なるフレームでは重み係数を１より大きな値として算出することを特徴とする。

【００４０】また、請求項１８記載の発明によれば、請求項１３から１７のいずれか１項に記載の音響モデル学習装置において、重み計算処理は、割り当てられた状態ごとに、算出した重み係数の和をそれぞれ算出し、算出した重み係数の和が、それぞれ等しい値となるように算出した重み係数を設定することを特徴とする。

【００４１】

【発明の実施の形態】（第１の実施形態）図１は、本発明の第１の実施形態における音響モデル学習装置の構成を示す図である。以下、図１を用いて、本実施形態における音響モデル学習装置の構成について説明する。なお、本実施形態では、音響モデルとして連続混合確率分布によるＨＭＭを用いる。上記の音響モデルでは、所定の音声からフレームごとに抽出された特徴量を示す確率分布を用いることによって、上記のフレームごとに分割された音声の断片が状態として表現され、その状態が構

成単位となる。

【００４２】音響モデル学習装置は、音声分析部１０１と、辞書部１０２と、フォワード・バックワード計算部１０３と、再評価部１０４と、ビタビ計算部１０５と、重み計算部１０６と、を有する。以下、図１を用いて音響モデル学習装置の各部位について説明する。

【００４３】音声分析部１０１には、音響モデルの学習に用いられる音声情報である学習用音声が入力される。なお、上記の学習用音声は、ビタビ計算部１０５にも入力される。

【００４４】音声分析部１０１は、入力された学習用音声を所定周期ごとに区切り、その区間を「フレーム」として、フレームごとに学習用音声の周波数分析を行う。上記の分析の結果抽出されたフレームごとの学習用音声の（音響的）特徴量は、フォワード・バックワード計算部１０３および再評価部１０４に入力される。なお、特徴量としては、音声のパワーを用いてもよいし、パワーの変化量、ケプストラム、あるいはケプストラム変化量等を用いてもよい。

【００４５】辞書部１０２には、音響モデルおよび正解列が入力される。上記の正解列は、所定の入力手段（図示せず）により入力される文字列の情報としてもよい。所定の入力手段は、音声分析部１０１およびビタビ計算部１０５に入力された学習用音声の内容を示す文字情報を正解列として辞書部１０２に入力する。

【００４６】また、辞書部１０２は、入力された音響モデル（以下、入力音響モデル）と入力された正解列とに基づいて、サブワードモデルによる学習用辞書を作成し、格納する。なお、サブワードモデルによる学習用辞書とは、入力された正解列（例えば、実際に存在する単語等）を、音素あるいは音節単位（サブワード単位）等に分割した状態列の情報である。また、辞書部１０２は、学習用辞書とは別に、任意の文字列の情報である「任意の文字列を表す辞書」を予め格納している。

【００４７】フォワード・バックワード計算部１０３は、辞書部１０２に格納されている学習用辞書を参照し、音声分析部１０１により抽出された学習用音声の特徴量と、入力された入力音響モデルと、に基づいて、フォワード・バックワード法によるフォワード確率とバックワード確率とを算出する。さらに、フォワード・バックワード計算部１０３は、算出したフォワード確率とバックワード確率とに基づいて、学習用音声の特徴量と入力音響モデルの状態との間の対応確率を算出する。フォワード・バックワード計算部１０３は、算出した対応確率を再評価部１０４へ出力する。

【００４８】フォワード・バックワード計算部１０３は、入力された学習用音声から変換されたフレームｔごとの特徴量をO_t（ｔは１以上Ｔ以下の整数）として、フォワード確率αを、以下に示す（式１．１）および（式１．２）に基づいて算出する。また、フォワード・バックワード計算部１０３は、バックワード確率βを、（式２．１）および（式２．２）により示されている式に基づいて算出する。

【００４９】また、フォワード・バックワード計算部１０３は、算出したフォワード確率αとバックワード確率βとを用いて、対応確率γを、（式３．１）により示される式に基づいて算出する。

【００５０】ビタビ計算部１０５には、音声分析部１０１と同様の学習用音声が入力される。また、ビタビ計算部１０５には、辞書部１０２を介して入力音響モデルが入力される。

【００５１】ビタビ計算部１０５は、入力された学習用音声を所定時間（フレーム）ごとに分割する。次に、ビタビ計算部１０５は、所定の文字情報を参照して、上記の各フレームに入力音響モデルに基づく状態あるいは複数の状態からなる状態列を最尤に割り当て、ビタビマッチング（Ｖｉｔｅｒｂｉ　Ｍａｔｃｈｉｎｇ）を行い、所定の最尤状態列を作成する。

【００５２】重み計算部１０６は、ビタビ計算部１０５により複数種類の所定の文字情報を参照して作成された複数種類の最尤状態列に基づいて重み係数R_tを算出する。

【００５３】再評価部１０４は、重み計算部１０６により算出された重み係数R_tと、フォワード・バックワード計算部１０３により算出された対応確率と、音声分析部１０１により抽出された特徴量と、フォワード・バックワード計算部１０３を介して入力された入力音響モデルと、に基づいて、音響モデルの各状態の統計量（混合重み、平均ベクトル、および共分散行列の各平均）を計算する。再評価部１０４は、抽出された統計量に基づいて、入力音響モデルの各パラメータ（混合重み、平均ベクトル、および共分散行列の各平均）を再評価する。再評価部１０４は、入力音響モデルの各パラメータの再評価に基づいて、音響モデルを作成する。再評価部１０４は、作成した音響モデルを、出力音響モデルとして出力する。

【００５４】再評価部１０４は、対応確率γに重み係数R_tを積算して重み付けを行う。再評価部１０４は、重み付けされた対応確率$\gamma \cdot R_t$を用いて、混合重みc_{jk}、平均ベクトルμ（ｔ，ｊ，ｋ）、および共分散行列Ｕ（ｊ，ｋ）の各平均を統計量として算出する。上記の統計量は、以下に示す（式５．１）、（式５．２）、および（式５．３）により与えられる。

【００５５】

【数５】

$$\overline{c}_{jk} = \frac{\sum_{t=1}^{T}{}_{s.t.\,Ot=Vk} \gamma(t,j,k) \cdot Rt}{\sum_{t=1}^{T} \sum_{k=1}^{M}{}_{s.t.\,Ot=Vk} \gamma(t,j,k) \cdot Rt} \quad \cdots \quad (式5.1)$$

$$\overline{\mu}(t,j,k) = \frac{\sum_{t=1}^{T}{}_{s.t.\,Ot=Vk} \gamma(t,j,k) \cdot Ot \cdot Rt}{\sum_{t=1}^{T}{}_{s.t.\,Ot=Vk} \gamma(t,j,k) \cdot Rt} \quad \cdots \quad (式5.2)$$

$$\overline{U}(j,k) = \frac{\sum_{t=1}^{T}{}_{s.t.\,Ot=Vk} \gamma(t,j,k) \cdot (Ot-\mu_{jk})(Ot-\mu_{jk})' \cdot Rt}{\sum_{t=1}^{T}{}_{s.t.\,Ot=Vk} \gamma(t,j,k) \cdot Rt} \quad \cdots \quad (式5.3)$$

【００５６】なお、混合重みc_{jk}は、ＨＭＭにおける状態S_jのｋ番目の混合分布要素に対する混合重みである。また、平均ベクトルμ（ｔ，ｊ，ｋ）は、ＨＭＭにおける状態S_jのｋ番目の混合分布要素の平均ベクトルである。また、共分散行列Ｕ（ｊ，ｋ）は、ＨＭＭにおける状態S_jのｋ番目の混合分布要素の共分散行列である。また、V_kは、文字列Ｖにおける所定の文字を示す。また、$(O_t-\mu_{jk})'$は、ベクトル$(O_t-\mu_{jk})$の対置ベクトルを表す。

【００５７】図２は、本発明の第１の実施形態における入力音響モデルが表現可能な音素セットを示す図である。また、図３は、本発明の第１の実施形態における音響モデル学習装置が作成する学習用辞書を示す図である。また、図４は、本発明の第１の実施形態における重み係数R_tを示す図である。また、図９は、本発明の第１の実施形態における音響モデル学習装置の動作の流れを示すフローチャートである。以下、図１～４を用い、図９に沿って本実施形態における音響モデル学習装置の動作について説明する。

【００５８】本実施形態では、学習用音声の一例として、所定の話者による「加藤今太郎（かとうこんたろう）」の発声を用いる。また、本実施形態では、入力音響モデル（初期モデル）として、上記の所定の話者による「かとうこんたろう」の発声を、「さとうこんたろう」と認識する音響モデルが与えられたとする。

【００５９】なお、ＨＭＭでは、１状態に対応する音声の長さは可変であり、ビタビマッチング等を用いることにより、ＨＭＭにおける最尤な状態系列が得られる。しかしながら、本実施形態では、簡単のために、入力音声は１４フレームの音声であり、１フレームにつき１状態が割り当てられているものとする。

【００６０】まず、所定の制御手段（図示せず）は、学習用音声が音声分析部１０１に入力されたか否かを判断する（ステップＳ９０１）。学習用音声が音声分析部１０１に入力されていないと判断された場合（ステップＳ９０１／Ｎｏ）、ステップＳ９０１の工程が繰り返される。

【００６１】学習用音声が音声分析部１０１に入力されたと判断された場合（ステップＳ９０１／Ｙｅｓ）、音声分析部１０１は、フレームごとに学習用音声の周波数を分析し、その分析した学習用音声の周波数に基づいて学習用音声の特徴量を抽出する（ステップＳ９０２）。抽出した学習用音声の特徴量は、フォワード・バックワード計算部１０３および再評価部１０４へ出力される。

【００６２】次に、所定の制御手段は、正解列および入力音響モデルが辞書部１０２に入力されたか否かを判断する（ステップＳ９０３）。正解列および入力音響モデルが入力されていないと判断された場合（ステップＳ９０３／Ｎｏ）、ステップＳ９０３の工程が繰り返される。

【００６３】正解列および入力音響モデルが辞書部１０２に入力されたと判断された場合（ステップＳ９０３／Ｙｅｓ）、辞書部１０２は、入力された正解列と入力音響モデルとに基づいて学習用辞書を作成し、作成した学習用辞書を格納する（ステップＳ９０４）。

【００６４】ここで、図２および図３を用いて、辞書部１０２が学習用辞書を作成する工程について説明する。図２には、本実施形態における入力音響モデルが表現で

きる音素の列（音素セット）が示されている。上記の音素セットは、入力音響モデルに含まれている。辞書部１０２は、上記の音素セットを用いて、学習用音声「かとうこんたろう」を「ｋ－ａ－ｔ－ｏ－ｕ－ｋ－ｏ－ｎｇ－ｔ－ａ－ｒ－ｏ－ｕ」と音素単位に分割する。分割した音素を、状態S_i（ｉは１以上１３以下の整数）にそれぞれ割り当て、図３に示されるような状態列、すなわち学習用音声に対応する学習用辞書を作成する。辞書部１０２は、作成した学習用辞書を格納する。

【００６５】辞書部１０２による学習用辞書作成後、フォワード・バックワード計算部１０３は、辞書部１０２により作成された学習用辞書を参照し、音声分析部１０１により抽出された特徴量に基づいて、フォワード確率およびバックワード確率を算出する（ステップＳ９０５）。

【００６６】次に、フォワード・バックワード計算部１０３は、算出したフォワード確率とバックワード確率とに基づいて対応確率を算出する（ステップＳ９０６）。

【００６７】所定の制御手段は、音声分析部１０１に入力された学習用音声と同様の学習用音声がビタビ計算部１０５に入力されたか否かを判断する。また、所定の制御手段は、入力音響モデルが辞書部１０２を介してビタビ計算部１０５に入力されたか否かを判断する（ステップＳ９０７）。学習用音声および入力音響モデルがビタビ計算部１０５に入力されていないと判断された場合（ステップＳ９０７／Ｎｏ）、ステップＳ９０７の工程が繰り返される。

【００６８】学習用音声および入力音響モデルがビタビ計算部１０５に入力されたと判断された場合（ステップＳ９０７／Ｙｅｓ）、ビタビ計算部１０５は、入力された学習用音声および入力音響モデルを用い、辞書部１０２により作成された学習用辞書を参照して、ビタビマッチングにより最尤状態列を生成する（ステップＳ９０８）。なお、学習用辞書を参照して生成された上記の最尤状態列を第１の最尤状態列とする。

【００６９】さらに、ビタビ計算部１０５は、入力された学習用音声および入力音響モデルを用い、辞書部１０２に格納されている任意の文字列を表す辞書を参照して、ビタビマッチングにより最尤状態列を生成する（ステップＳ９０９）。なお、任意の文字を表す辞書を参照して生成された上記の最尤状態列を第２の最尤状態列とする。

【００７０】次に、重み計算部１０６は、ビタビ計算部１０５により生成された第１の最尤状態列の各状態と第２の最尤状態列の各状態を比較し、以下に示す（式６．１）および（式６．２）により与えられる重み係数R_tを算出する（ステップＳ９１０）。なお、重み係数R_tは、学習用音声の各フレームにそれぞれ対応するように算出される。

【００７１】

【数６】

$$R_t = 0 \quad (\exists j \ s.t. \ \sum_k \gamma(t,j,k) \neq \sum_k \gamma(t,j,k)) \quad \cdots \quad (式6.1)$$

$$R_t = 1 \quad (\forall j, \ \sum_k \gamma(t,j,k) = \sum_k \gamma(t,j,k)) \quad \cdots \quad (式6.2)$$

【００７２】話者による誤発声あるいは品質の低い音声を学習用音声として用いた場合、入力された正解列と入力音響モデルにより認識される学習用音声との間で差異が発生する可能性、つまり、入力された学習用音声による所定の言語単位（例えば、音素単位、音節単位等）の音声サンプルが音響モデルにより誤認識される可能性が高い。上記の誤認識された音声サンプルが出力音響モデルに大きく反映しないようすることによって、信頼性の高い出力音響モデルを得ることが可能となる。

【００７３】重み計算部１０６は、第１の最尤状態列における各状態と、第２の最尤状態列における各状態と、をフレームごとに比較し、上記の（式６．１）および（式６．２）に基づいて重み係数R_tを算出する。

【００７４】（式６．１）は、所定のフレームにおいて、第１の最尤状態列と第２の最尤状態列との間に差異が発生した場合の重み係数R_tを与える式であり、上記の場合、重み係数R_tは「０」として算出される。

【００７５】（式６．２）は、全てのフレームにおいて、第１の最尤状態列と第２の最尤状態列とが一致した場合の重み係数R_tを与える式であり、上記の場合、重み係数R_tは「１」として算出される。

【００７６】話者の誤発声等により学習用音声の品質が低下した場合、その品質低下が生じた部分に対応するフレームに割り当てられている第１の最尤状態列の状態と、第２の最尤状態列の状態との間に差異が発生する。従って、信頼性の高い出力音響モデルを得るためには、上記の差異が生じた部分が出力音響モデルに反映されないようにする必要がある。

【００７７】本実施形態では、学習用音声における高品質部分（所定のフレームにおいて第１の最尤状態列の状態と第２の最尤状態列とが一致した状態）の重み係数R_tを「１」とし、低品質部分の重み係数R_tを高品質部分の重み係数R_tよりも低い値である「０」とすることによって、学習用音声の低品質部分、すなわち学習用音声が入力音響モデルにより誤認識されている部分が出力音響モデルに反映されないようにしている。

【００７８】本実施形態における入力音響モデルでは、学習用音声「かとうこんたろう」は、「さとうこんたろ

う」と認識される。上記のような場合、「か」の部分が実際にどのような発声であったか不明であるが、「か」の部分における音素「ｋ」が入力音響モデルにより誤認識されている。音素「ｋ」のモデルが正しく認識される出力音響モデルを作成するためには、「か」の部分の音素「ｋ」が出力音響モデルに反映しないように設定される必要がある。

【００７９】図１４は、（式６．１）および（式６．２）により図３の学習用辞書に与えられる重み係数R_tを示す図である。R_t（ｔ＝１～１３）は、それぞれS_i（ｉ＝１～１３）における重み係数である。図１４に示されているように、「か」の部分の音素「ｋ」（＝S_1）における重み係数R_1を「０」とし、他の音素（S_2～S_{13}）における重み係数R_2～R_{13}を「１」とすることによって、「か」の部分の音素「ｋ」が出力音響モデルに反映しないようにすることが可能となり、信頼性の高い音響モデルを作成することが可能となる。

【００８０】なお、本実施形態では、重み係数R_1を「０」とすることにより、「か」の部分の音素「ｋ」が出力音響モデルに反映しないようにしたが、重み係数R_1を「０以上１未満の任意の値」に設定することによって、「か」の部分の音素「ｋ」が出力音響モデルに与える影響を調整することが可能となる。

【００８１】以下、再び図９のフローチャートに沿って音響モデル学習装置の動作について説明を進める。再評価部１０４は、重み計算部１０６により算出された重み係数R_tと、音声分析部１０１により抽出された特徴量と、フォワード・バックワード計算部１０３により算出された対応確率と、に基づいて、音響モデルの各統計量（混合重み、平均ベクトル、および共分散行列の各平均）を算出する（ステップＳ９１１）。

【００８２】再評価部１０４は、音響モデルの各統計量算出後、算出した統計量に基づいて、フォワード・バックワード計算部１０３を介して入力された入力音響モデルの各パラメータ（混合重み分布、平均ベクトル、および共分散行列の各平均）を再評価し、出力音響モデルを作成する（ステップＳ９１２）。作成された出力音響モデルは、再評価部１０４から出力される（ステップＳ９１３）。出力音響モデル出力後、音響モデル学習装置は、動作を終了する。

【００８３】（第２の実施形態）以下、特記しない限り、本発明の第２の実施形態における音響モデル学習装置の構成および動作は、本発明の第１の実施形態における音響モデル学習装置の構成および動作と同様であるとする。

【００８４】一般に、騒音環境が学習用音声の品質を低下させる場合、学習用音声の誤認識は、単一の音素にとどまらず、その音素の周辺音素にも影響を与える。第１の実施形態では、重み係数R_tを音素ごとに設定していたが、環境騒音などの理由により複数の音素にわたって誤認識される場合、音節単位で重み付けを行うことによって、より信頼性の高い出力音響モデルを作成することが可能となる。

【００８５】図５は、本発明の第２の実施形態における重み係数R_tを示す図である。第１の実施形態と同様に重み係数R_t（ｔ＝１～１３）は、それぞれ図３における状態S_i（ｉ＝１～１３）に対応する。

【００８６】第１の実施形態では、「か」の音素「ｋ」（＝S_1）の重み係数R_1を「０」に設定していた。本実施形態では、学習用音声「かとうこんたろう」における音節「か（ｋ－ａ）」において、品質が低下し、第１の最尤状態列と第２の最尤状態列との間に差異が生じている。上記のように音節単位で学習用音声の品質低下が生じている場合、「か」の音素「ｋ」（＝S_1）の重み係数R_1と、音素「ａ」（＝S_2）の重み係数R_2と、をそれぞれ「０」に設定することによって、音素「ｋ」（＝S_1）の重み係数R_1のみを「０」とする場合と比較して、より信頼性の高い出力音響モデルを作成することが可能となる。

【００８７】なお、本実施形態では、重み係数R_1およびR_2を「０」とすることにより、「か」の部分の音素「ｋ」および音素「ａ」が出力音響モデルに反映しないようにしたが、重み係数R_1およびR_2を「０以上１未満の任意の値」に設定することによって、「か」の部分の音素「ｋ」および音素「ａ」が出力音響モデルに与える影響を調整することが可能となる。

【００８８】（第３の実施形態）以下、特記しない限り、本発明の第３の実施形態における音響モデル学習装置の構成および動作は、本発明の第１の実施形態における音響モデル学習装置の構成および動作と同様であるとする。

【００８９】第２の実施形態では、騒音環境による学習用音声の誤認識は、単一の音素にとどまらず、その音素の周辺音素にも影響を与える場合について説明した。第２の実施形態では、重み係数R_tを音節ごとに設定していたが、環境騒音などの理由により誤認識される音素の範囲が音節単位よりもさらに広い範囲にわたって存在する場合、重み付けする音素の範囲を音節単位よりもさらに拡大し、単語単位とすることによって、より信頼性の高い出力音響モデルを作成することが可能となる。

【００９０】図６は、本発明の第３の実施形態における重み係数R_tを示す図である。第１の実施形態と同様に重み係数R_t（ｔ＝１～１３）は、それぞれ図３における状態S_i（ｉ＝１～１３）に対応する。

【００９１】第１の実施形態では、「か」の音素「ｋ」（＝S_1）の重み係数R_1を「０」に設定していた。また、第２の実施形態では、「か」の音素「ｋ」（＝S_1）の重み係数R_1と、音素「ａ」（＝S_2）の重み係数R_2と、をそれぞれ「０」に設定していた。本実施形態では、学習用音声「かとうこんたろう」における単

語「かとう（ｋ－ａ－ｔ－ｏ－ｕ）」において、品質が低下し、第１の最尤状態列と第２の最尤状態列との間に差異が生じている。上記のように単語単位で学習用音声の品質低下が生じている場合、単語「かとう（ｋ－ａ－ｔ－ｏ－ｕ）」における音素「ｋ」（＝S_1）、音素「ａ」（＝S_2）、音素「ｔ」（＝S_3）、音素「ｏ」（＝S_4）、および音素「ｕ」（＝S_5）それぞれに対応する重み係数R_1～R_5を「０」とすることによって、音素単位あるいは音節単位で重み係数R_tを「０」とする場合と比較して、より信頼性の高い出力音響モデルを作成することが可能となる。

【００９２】なお、本実施形態では、重み係数R_1～R_5を「０」とすることにより、「かとう」の部分の音素「ｋ－ａ－ｔ－ｏ－ｕ」が出力音響モデルに反映しないようにしたが、重み係数R_1～R_5を０以上１未満の任意の値に設定することによって、「かとう」の部分の音素「ｋ－ａ－ｔ－ｏ－ｕ」が出力音響モデルに与える影響を調整することが可能となる。

【００９３】（第４の実施形態）以下、特記しない限り、本発明の第４の実施形態における音響モデル学習装置の構成および動作は、本発明の第１の実施形態における音響モデル学習装置の構成および動作と同様であるとする。

【００９４】上記の第１から第３の実施形態では、第１の最尤状態列と第２の最尤状態列との間で差異が生じた部分（学習用音声の品質が低下した部分）の重み係数R_tを「０」に設定し、出力音響モデルに反映されないようにしていた。本実施形態における音響モデル学習装置は、学習用音声における誤発声あるいは品質の低い音声が生じた部分を発声の一変化として積極的に取り入れ、学習用音声の高品質部分よりも高い重み係数R_tを設定することによって、低品質の学習用音声のサンプル数を増加させ、低品質の学習用音声に対する認識性能を向上させる。

【００９５】図７は、本発明の第４の実施形態における重み係数R_tを示す図である。図７に示される重み係数R_tは、以下に示す（式７．１）および（式７．２）により与えられる。

【００９６】

【数７】

$$R_t=10 \quad (\exists j,\ s.t.\ \sum_k \gamma(t,j,k) \neq \sum_k \gamma(t,j,k)) \quad \cdots \quad (式7.1)$$

$$R_t=1 \quad (\forall j,\ \sum_k \gamma(t,j,k) = \sum_k \gamma(t,j,k)) \quad \cdots \quad (式7.2)$$

【００９７】本実施形態では、第１の実施形態と同様に、所定の話者により入力された「かとうこんたろう」という学習用音声を、「さとうこんたろう」と認識する音響モデルが入力される。第１の実施形態では、「か」の音素「ｋ」（＝S_1）に対応する重み係数R_1を「０」に設定し、出力音響モデルに反映しないようにすることによって、信頼性の高い出力音響モデルを作成していた。本実施形態では、第１の最尤状態列と第２の最尤状態列との間で差異が発生した「か」の音素「ｋ」（＝S_1）に、第１の最尤状態列と第２の最尤状態列との間で一致した他の音素に設定された「重み係数R_t＝１（ｔ＝2～１３）」よりも高い「重み係数R_1＝１０」を設定する。

【００９８】上記のように、「重み係数R_1＝１０」と設定することによって、十分に学習されていない稀な特徴と考えられる「か」の音素「ｋ」（＝S_1）を、他の音素よりも出力音響モデルに大きく反映させることが可能となる。

【００９９】なお、本実施形態では、重み係数R_tによる重み付けを音素単位で行ったが、第２の実施形態のように音節単位で行ってもよいし、第３の実施形態のように単語単位で行ってもよい。

【０１００】また、本実施形態では、正解列と入力音響モデルにより認識された学習用音声との間で差異が生じた音素に対応する重み係数R_tを「１０」としたが、正解列と学習用音声との間で一致した音素と比較して大きな数値であれば、差異が生じた音素に対応する重み係数R_tは、他の値であってもよい。

【０１０１】（第５の実施形態）以下、特記しない限り、本発明の第５の実施形態における音響モデル学習装置の構成および動作は、本発明の第１の実施形態における音響モデル学習装置の構成および動作と同様であるとする。

【０１０２】統計モデルの信頼性は、統計モデルのパラメータ学習に用いられた音声サンプル（音素、音節、あるいは単語）の量により大きく影響される。従って、各音響モデルにおける信頼性を均一化するためには、入力される各音声サンプルの量に著しい偏りが生じないようにする必要がある。

【０１０３】本実施形態では、第１の最尤状態列における各状態ごとの重み係数R_tの和を一定にし、入力される所定の言語単位（音素、音節、あるいは単語等）の各音声サンプルにおけるサンプル数を均一化する。

【０１０４】図１０は、本発明の第５の実施形態における音響モデル学習装置の動作の流れを示すフローチャートである。以下、図１を用い、図１０に沿って、本実施形態における音響モデル学習装置の動作について説明する。

【０１０５】本実施形態では、第１の実施形態と同様に、学習用音声の一例として、所定の話者による「加藤

今太郎（かとうこんたろう）」の発声を用いる。

【0106】まず、所定の制御手段（図示せず）は、学習用音声が音声分析部101に入力されたか否かを判断する（ステップS1001）。学習用音声が音声分析部101に入力されていないと判断された場合（ステップS1001／No）、ステップS1001の工程が繰り返される。

【0107】学習用音声が音声分析部101に入力されたと判断された場合（ステップS1001／Yes）、音声分析部101は、フレームごとに学習用音声の周波数を分析し、その分析した学習用音声の周波数に基づいて学習用音声の特徴量を抽出する（ステップS1002）。抽出した学習用音声の特徴量は、フォワード・バックワード計算部103および再評価部104へ出力される。

【0108】次に、所定の制御手段は、正解列および入力音響モデルが辞書部102に入力されたか否かを判断する（ステップS1003）。正解列および入力音響モデルが入力されていないと判断された場合（ステップS1003／No）、ステップS1003の工程が繰り返される。

【0109】正解列および入力音響モデルが辞書部102に入力されたと判断された場合（ステップS1003／Yes）、辞書部102は、入力された正解列と入力音響モデルとに基づいて学習用辞書を作成し、作成した学習用辞書を格納する（ステップS1004）。

【0110】辞書部102による学習用辞書作成後、フォワード・バックワード計算部103は、辞書部102により作成された学習用辞書を参照し、音声分析部101により抽出された特徴量に基づいて、フォワード確率およびバックワード確率を算出する（ステップS1005）。

$$R_t = 1 \quad \left(\sum_{\substack{n \le t \\ O_t \le V_n}} R_n \le M\right) \quad \cdots \quad \text{(式9.1)}$$

$$R_t = 0 \quad \left(\sum_{\substack{n \le t \\ O_t \le V_n}} R_n > M\right) \quad \cdots \quad \text{(式9.2)}$$

$$\text{s.t. } M = \min_{j} \sum_{t} \sum_{k} \gamma(t, j, k) \quad \cdots \quad \text{(式9.3)}$$
$$O_t = V_n$$

【0111】次に、フォワード・バックワード計算部103は、算出したフォワード確率とバックワード確率とに基づいて対応確率を算出する（ステップS1006）。

【0112】所定の制御手段は、音声分析部101に入力された学習用音声と同様の学習用音声がビタビ計算部105に入力されたか否かを判断する。また、所定の制御手段は、入力音響モデルが辞書部102を介してビタビ計算部105に入力されたか否かを判断する（ステップS1007）。学習用音声および入力音響モデルがビタビ計算部105に入力されていないと判断された場合（ステップS1007／No）、ステップS1007の工程が繰り返される。

【0113】学習用音声および入力音響モデルがビタビ計算部105に入力されたと判断された場合（ステップS1007／Yes）、ビタビ計算部105は、入力された学習用音声および入力音響モデルを用い、辞書部102により作成された学習用辞書を参照して、ビタビマッチングにより最尤状態列を生成する（ステップS1008）。なお、学習用辞書を参照して生成された上記の最尤状態列を第1の最尤状態列とする。

【0114】次に、重み計算部106は、ビタビ計算部105により生成された第1の最尤状態列の各状態を参照し、以下に示す（式8．1）、（式9．1）、（式9．2）、および（式9．3）に基づいて、重み係数R_tを算出する（ステップS1009）。

【0115】

【数8】

$$\forall i, j, \quad \sum_{O_n = V_i} R_n = \sum_{O_n = V_j} R_n \quad \cdots \quad \text{(式8.1)}$$

【0116】

【数9】

【0117】本実施形態では、上記の（式8．1）で与えられる条件により、学習用音声を構成する同一の音声サンプル（音素、音節、あるいは単語単位）が割り当てられている状態ごとに重み係数R_tの和をとり、重み係数R_tの和が等しくなるように、重み係数R_tを算出することによって、各音声サンプルがそれぞれ出力音響モデルに与える影響が均一になる。

【0118】本実施形態では、本発明の第1の実施形態と同様に図3に示される学習用辞書が生成されるとする。図8は、本発明の第5の実施形態における重み係数R_tを示す図である。図8に示される重み係数R_tは、上記の（式9．1）、（式9．2）および（式9．3）に基づいて設定されている。なお、図8における重み係数R_t（t＝1～13）は、図3に示されている状態S_i（i＝1～13）にそれぞれ対応している。

【0119】本実施形態では、割り当てられたフレーム

の値が小さなものから順に、学習用音声を構成する音素を観測した場合、初めて観測された種類の音素に対応する重み係数R_tを「１」とし、以前観測された種類の音素に対応する重み係数R_tを「０」としている。

【０１２０】以下、図３および図８を用いて説明すると、例えば、S_6の音素「ｋ」は、すでにS_1において観測されているので重み係数R_6は「０」に設定されている。一方、S_{11}の音素「ｒ」は、S_1～S_{10}において観測されていないので重み係数R_{11}は「１」に設定されている。

【０１２１】上記のように重み係数R_tが算出されることによって、同一種類の音素に付加されている重み係数R_tの和は、それぞれ「１」となり、各音素が音声サンプルとして収集される回数が均等となる。

【０１２２】以下、再び図１０のフローチャートに沿って音響モデル学習装置の動作について説明を進める。再評価部１０４は、重み計算部１０６により算出された重み係数R_tと、音声分析部１０１により抽出された特徴量と、フォワード・バックワード計算部１０３により算出された対応確率と、に基づいて、音響モデルの各統計量（混合重み、平均ベクトル、および共分散行列の各平均）を算出する（ステップＳ１０１０）。

【０１２３】再評価部１０４は、音響モデルの各統計量算出後、算出した統計量に基づいて、フォワード・バックワード計算部１０３を介して入力された入力音響モデルの各パラメータ（混合重み分布、平均ベクトル、および共分散行列の各平均）を再評価し、出力音響モデルを作成する（ステップＳ１０１１）。作成された出力音響モデルは、再評価部１０４から出力される（ステップＳ１０１２）。出力音響モデル出力後、音響モデル学習装置は、動作を終了する。

【０１２４】本実施形態では、以上説明したように、同一の音声サンプル（音素、音節、あるいは単語）が割り当てられた状態ごとの重み係数R_tの和を一定とすることによって、各音声サンプル（音素、音節、あるいは単語単位）のサンプル量および出力音響モデルに与える影響を均一化し、信頼性の高い出力音響モデルを作成することを可能としている。

【０１２５】また、音響モデル学習装置は、入力される学習用音声からフレームごとに特徴量を抽出する音声分析処理と、所定の音声からフレームごとに抽出された特徴量を示す確率分布を用いて、所定の音声におけるフレームごとの特徴量を状態として表現し、状態を構成単位とする入力音響モデルと、学習用音声の内容を示す文字列情報である正解列と、に基づいて、入力音響モデルにおける状態に正解列を割り当てた状態列の情報である学習用辞書を生成する辞書生成処理と、辞書生成処理により生成された学習用辞書を参照し、学習用音声の特徴量と入力音響モデルにおける状態との対応確率を学習用音声のフレームごとに算出する対応確率算出処理と、所定の文字列を用いて、入力音響モデルにより表現される状態あるいは複数の状態からなる状態列を、学習用音声のフレームごとに最尤に割り当て、所定の最尤状態列を生成する最尤状態列生成処理と、最尤状態列生成処理により生成された所定の最尤状態列に基づいて、対応確率に重み付けする際に付加する係数である重み係数を、学習用音声のフレームごとに算出する重み計算処理と、対応確率算出処理により算出された対応確率と、重み計算処理により算出された重み係数と、音声分析処理により算出された特徴量と、に基づいて統計量を算出し、算出した統計量に基づいて、入力音響モデルのパラメータを再推定し、出力音響モデルを作成する再評価処理と、を行う。上記の処理は、音響モデル学習装置が有するコンピュータプログラムにより実行されるが、上記のプログラムは、光ディスクあるいは磁気ディスク等の記録媒体に記録され、上記の記録媒体からロードされるようにしてもよい。

【０１２６】なお、上記の実施形態は本発明の好適な実施の一例であり、本発明の実施形態は、これに限定されるものではなく、本発明の要旨を逸脱しない範囲において種々変形して実施することが可能となる。

【０１２７】

【発明の効果】以上説明したように、本発明は、学習用音声のフレームごとに重み係数を算出し、上記の重み係数による重み付けを出力音響モデルに反映させることによって、観測された音声サンプルのうち音響モデルの作成に有用なものだけを抽出し、信頼性の高い音響モデルを作成することが可能となる。

【０１２８】また、本発明は、品質の高い所定の言語単位（音素、音節、あるいは単語等）の音声サンプルの重み付け係数を「１」とし、品質の低い音声サンプルの重み付け係数を「０」とすることによって、品質の低い音声サンプルが出力音響モデルに反映しないようにすることが可能となる。

【０１２９】また、本発明は、品質の高い所定の言語単位の音声サンプルの重み付け係数を「１」とし、品質の低い音声サンプルの重み付け係数を「１より大きな任意の値」とすることによって、品質の低い音声サンプルに対する音声認識の精度が高い出力音響モデルを作成することが可能となる。

【０１３０】また、本発明は、同一の音声サンプル（音素、音節、あるいは単語）が割り当てられた状態ごとの重み係数の和を一定とすることによって、各音声サンプル（音素、音節、あるいは単語単位）のサンプル量および出力音響モデルに与える影響を均一化し、信頼性の高い出力音響モデルを作成することが可能となる。

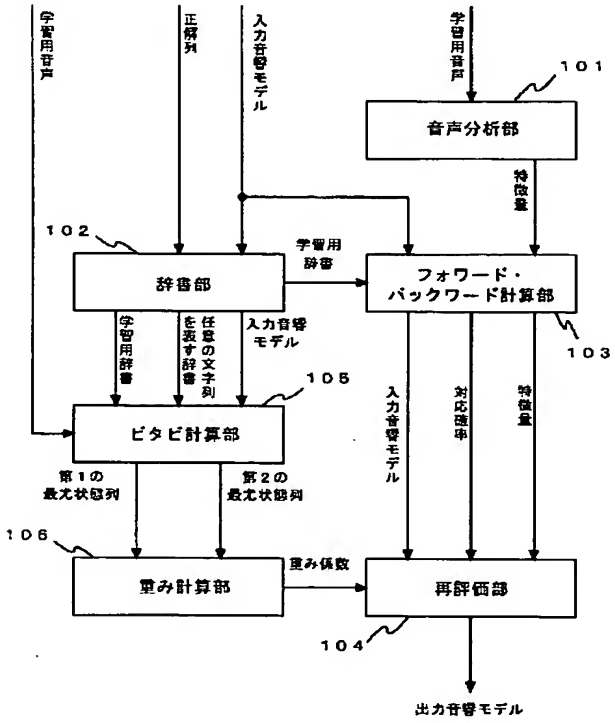
【図面の簡単な説明】

【図１】本発明の第１の実施形態における音響モデル学習装置の構成を示す図である。

【図２】本発明の第１の実施形態における入力音響モデ

ルが表現可能な音素セットを示す図である。
【図３】本発明の第１の実施形態における音響モデル学習装置が作成する学習用辞書を示す図である。
【図４】本発明の第１の実施形態における重み係数R_tを示す図である。
【図５】本発明の第２の実施形態における重み係数R_tを示す図である。
【図６】本発明の第３の実施形態における重み係数R_tを示す図である。
【図７】本発明の第４の実施形態における重み係数R_tを示す図である。
【図８】本発明の第５の実施形態における重み係数R_tを示す図である。
【図９】本発明の第１の実施形態における音響モデル学習装置の動作の流れを示すフローチャートである。
【図１０】本発明の第５の実施形態における音響モデル学習装置の動作の流れを示すフローチャートである。
【符号の説明】
１０１　音声分析部
１０２　辞書部
１０３　フォワード・バックワード計算部
１０４　再評価部
１０５　ビタビ計算部
１０６　重み計算部

【図１】



【図４】

R1	0
R2	1.0
R3	1.0
R4	1.0
R5	1.0
R6	1.0
R7	1.0
R8	1.0
R9	1.0
R10	1.0
R11	1.0
R12	1.0
R13	1.0

【図８】

R1	1.0
R2	1.0
R3	1.0
R4	1.0
R5	1.0
R6	0
R7	0
R8	1.0
R9	0
R10	0
R11	1.0
R12	0
R13	0

【図５】

R1	0
R2	0
R3	1.0
R4	1.0
R5	1.0
R6	1.0
R7	1.0
R8	1.0
R9	1.0
R10	1.0
R11	1.0
R12	1.0
R13	1.0

【図６】

R1	0
R2	0
R3	0
R4	0
R5	0
R6	1.0
R7	1.0
R8	1.0
R9	1.0
R10	1.0
R11	1.0
R12	1.0
R13	1.0

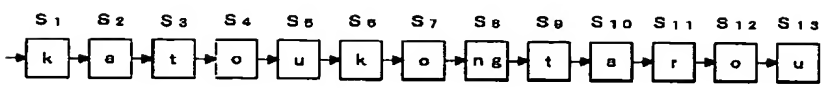
【図７】

R1	10.0
R2	1.0
R3	1.0
R4	1.0
R5	1.0
R6	1.0
R7	1.0
R8	1.0
R9	1.0
R10	1.0
R11	1.0
R12	1.0
R13	1.0

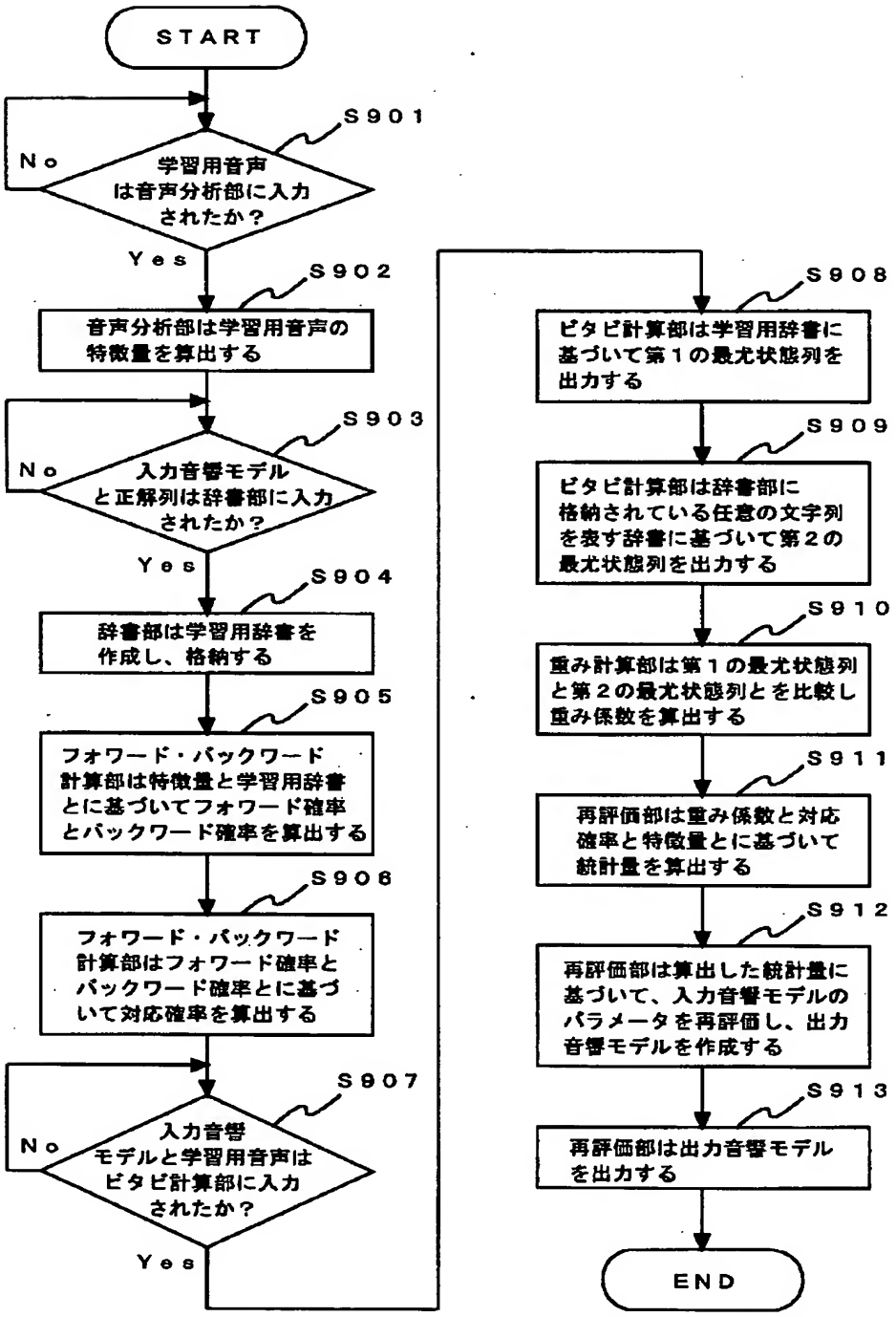
【図２】

a, b, ch, d, e, f, g, h, i, j, k, m, n, ng, o, p, r, s, sh, t, ts, u, v, w, y, z, zh

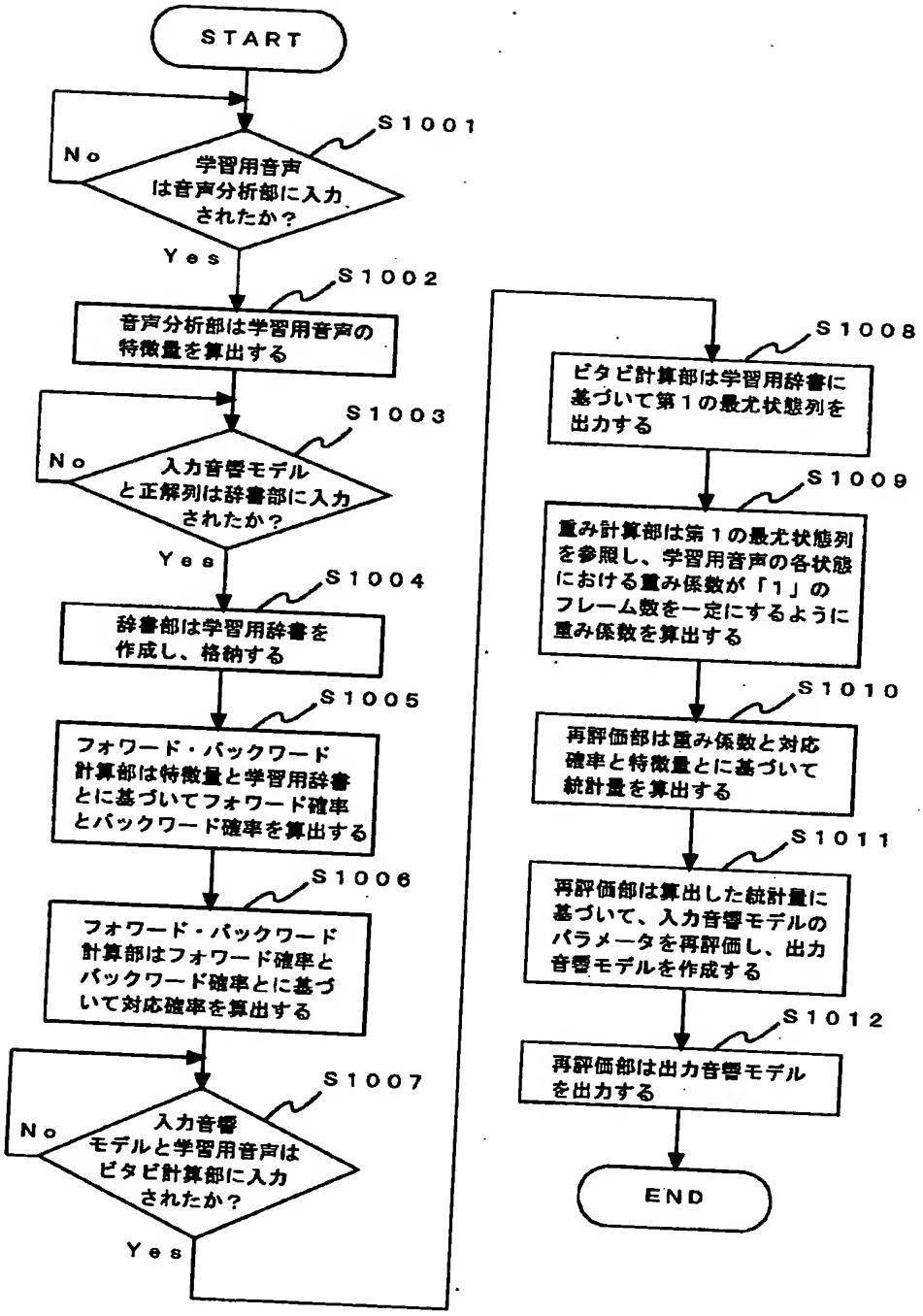
【図３】



【図9】


START
S901
No
学習用音声は音声分析部に入力されたか？
Yes
S902
音声分析部は学習用音声の特徴量を算出する
S903
No
入力音響モデルと正解列は辞書部に入力されたか？
Yes
S904
辞書部は学習用辞書を作成し、格納する
S905
フォワード・バックワード計算部は特徴量と学習用辞書とに基づいてフォワード確率とバックワード確率を算出する
S906
フォワード・バックワード計算部はフォワード確率とバックワード確率とに基づいて対応確率を算出する
S907
No
入力音響モデルと学習用音声はビタビ計算部に入力されたか？
Yes
S908
ビタビ計算部は学習用辞書に基づいて第１の最尤状態列を出力する
S909
ビタビ計算部は辞書部に格納されている任意の文字列を表す辞書に基づいて第２の最尤状態列を出力する
S910
重み計算部は第１の最尤状態列と第２の最尤状態列とを比較し重み係数を算出する
S911
再評価部は重み係数と対応確率と特徴量とに基づいて統計量を算出する
S912
再評価部は算出した統計量に基づいて、入力音響モデルのパラメータを再評価し、出力音響モデルを作成する
S913
再評価部は出力音響モデルを出力する
END

【図10】


START
S1001
No
学習用音声は音声分析部に入力されたか？
Yes
S1002
音声分析部は学習用音声の特徴量を算出する
S1003
No
入力音響モデルと正解列は辞書部に入力されたか？
Yes
S1004
辞書部は学習用辞書を作成し、格納する
S1005
フォワード・バックワード計算部は特徴量と学習用辞書とに基づいてフォワード確率とバックワード確率を算出する
S1006
フォワード・バックワード計算部はフォワード確率とバックワード確率とに基づいて対応確率を算出する
S1007
No
入力音響モデルと学習用音声はビタビ計算部に入力されたか？
Yes
S1008
ビタビ計算部は学習用辞書に基づいて第１の最尤状態列を出力する
S1009
重み計算部は第１の最尤状態列を参照し、学習用音声の各状態における重み係数が「１」のフレーム数を一定にするように重み係数を算出する
S1010
再評価部は重み係数と対応確率と特徴量とに基づいて統計量を算出する
S1011
再評価部は算出した統計量に基づいて、入力音響モデルのパラメータを再評価し、出力音響モデルを作成する
S1012
再評価部は出力音響モデルを出力する
END